

AD-A096 680

NAVAL POSTGRADUATE SCHOOL MONTEREY CA

F/6 12/1

STATISTICAL METHODS, SOME OLD, SOME NEW: A TUTORIAL SURVEY, (U)

JAN 81 D P GAVER

UNCLASSIFIED

NPS55GV-81-002

NL

1 OF 1  
AD-A096 680

END  
DATE  
FILMED  
4-81  
DTIC

AD A 096680

14

NPS55Gy-81-002

LEVEL

2

# NAVAL POSTGRADUATE SCHOOL

Monterey, California



DTIC  
MAR 23 1981

9) Technical report

6.

STATISTICAL METHODS, SOME OLD, SOME NEW:  
A TUTORIAL SURVEY,

by

1. Donald P. /Gaver

11 Jan 1981

12471

Approved for public release; distribution unlimited.

Prepared for:  
Office of Naval Research  
Arlington, VA 22217

17 R1110501  
6 R111751

231127

81 3 23 026

ENCLOSURE

NAVAL POSTGRADUATE SCHOOL  
Monterey, California

Rear Admiral J. J. Ekelund  
Superintendent

David A. Schradly  
Acting Provost

This work was supported in part by the Office of Naval  
Research, Alexandria, VA 22217.

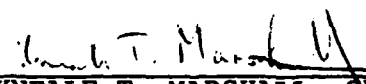
Reproduction of all or part of this report is authorized.


Prepared by:

  
DONALD P. GAVER, Professor  
Department of Operations Research

Reviewed by:

Released by:

  
KNEALE T. MARSHALL, Chairman  
Department of Operations Research

  
WILLIAM M. TOLLES  
Dean of Research

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55-81-002	2. GOVT ACCESSION NO. AD-A096680	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Statistical Methods, Some Old, Some New: A Tutorial Survey		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Donald P. Gaver		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61153N, RR014-05-01 NR# 042-411, Noo01481WR1000
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Alexandria, VA 22217		12. REPORT DATE January 1981
		13. NUMBER OF PAGES 35
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)  <div style="display: flex; justify-content: space-between;"> <div> statistics data analysis graphics </div> <div> transformations regression time series </div> </div>		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)  This report describes some new graphical and robust methods of statistical analysis. It also contains brief accounts of logistic (categorical) regression, and of regression in the presence of autocorrelated disturbances.		

DD FORM 1473  
1 JAN 73EDITION OF 1 NOV 65 IS OBSOLETE  
S/N 0102-014-6601UNCLASSIFIED 25/4  
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

STATISTICAL METHODS, SOME OLD, SOME NEW

A TUTORIAL SURVEY

Donald P. Gaver

This report is a revised and expanded version of material on statistics presented at the Summer School on Remote Sensing in Meteorology, Oceanography, and Hydrology, held at the University of Dundee, Scotland during the month of September 1980; its directors were Professor A. P. Cracknell and Dr. G. Ostrem.

The attendees, both students and faculty, were from many countries and professional and educational backgrounds. There were, for example, physicists, electrical engineers, atmospheric and geophysical scientists, physical geographers, photographers--and two statisticians: Dr. Ed Wegman of ONR, Washington, and Dr. Donald Gaver of the Naval Postgraduate School, Monterey, California. The present report was largely assembled by D.G., with inputs from E.W. A shortened version was placed on transparencies and formed the basis for two one-hour-plus lectures.

Those who attended these lectures appeared to be quite positive in their response. It will be noted that no attempt was made to present much formal mathematical material, perhaps to the listeners' surprise. In view of the background of most participants, this seemed appropriate. An attempt was made to lead the listeners into some of the approaches and concerns of the data analyst; particularly one who might be working with "environmental" data. In the course of talking to participants during the school, and particularly following these lectures, I discovered that there

was a good deal of interest in statistical methodology for use in the atmospheric and geophysical sciences. I hope to follow up on some of the contacts made, and perhaps engage in collaborative activities. There seem to be many who are interested, and much to be done in applying statistical thinking and methodology to the various, generally environmental, geophysical, or atmospheric areas of interest represented at this exciting and stimulating summer school.

# STATISTICAL METHODS, SOME OLD, SOME NEW

## A TUTORIAL SURVEY

Donald P. Gaver

### 1. Introduction

The purpose of the two lectures with the above title is to introduce to some, and review for others, selected topics in statistical methods. It is hoped that these topics will be useful to workers in remote sensing. Few details will be given, but references will be provided so that those interested can go further.

Statistics can be tentatively defined as the science, technology, and art of drawing quantitative inferences from data. The subject is always in a process of development, urged by the needs of those who have, or plan to obtain, data, and further stimulated by conceptual developments and the widening possibilities of using digital computers for storing, analyzing, displaying and finally understanding the meaning of that data.

#### 1.1. Phases of a Statistical Inquiry

It seems useful to distinguish several phases of a statistical inquiry

- o Objectives of the inquiry,
- o Data acquisition,
- o Data exploration,
- o Model choice or construction,
- o Confirmatory or validation analysis,
- o Communication of results.

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Avail and/or	
Dist	Special

*A*

Our term data acquisition subsumes the choice of what data to take in a particular problem setting, experimental design for economy and avoidance of bias, etc. This topic is crucially important, but for this audience requires a substantive knowledge of remote sensing issues that we do not yet possess. It will not be discussed here. By data exploration is meant the activity of examining data, both graphically and through numerical summaries, for the purpose of revealing properties of the data itself, and, with luck, of the process giving rise to that data. John Tukey, e.g. (1977) has shown us that there is much to value and learn about exploration of data before, or even without, the use of the probability theory structure that has traditionally seemed so necessary for formal statistics. We will discuss several of these simple exploratory approaches in the first lecture. Model choice or construction is the phase of inquiry that brings to bear subject matter knowledge to "explain" data behavior. Here probability theory may well enter to represent measurement errors, or fluctuations in natural phenomena, such as the heights of wind-driven water waves. The confirmatory analysis phase is concerned with developing quantitative expressions for uncertainties inherent in simple summary statements, and in more complex characterizations and predictions made from formal models. This is the formal inference phase to which mathematical (especially probability-theoretic) methods continue to contribute.



## 2. Simple Graphics and Summaries of Data

### 2.1. Stem and Leaf (Successor to Histogram); Transformation

Suppose we have a batch of measurements on some physical quantity. We wish to display these numbers in frequency-of-occurrence style, losing as little detail as possible.

Here is such a batch:

0.92, 1.14, 2.00, 2.66, 6.52, 4.95, 2.76, 2.13, 1.14, 9.72  
11.24, 0.25, 3.31, 12.66, 0.46, 2.77, 0.35, 66.67, 0.59, 1.84

To get a stem-and-leaf display draw a vertical line (stem) and decide on reasonable class intervals: first try intervals  $0 \rightarrow 10-$ ,  $10 \rightarrow 20-$ ,  $20 \rightarrow 30-$ , etc. Now put the integer part of the number by convention (rounded to the nearest integer) to the right of the stem, e.g. 0.92 becomes 1, 1.14 becomes 1, ...6.52 becomes 7, etc. The stem and leaf appears as follows:

0		1123532103030127
1		013
2		
3		
4		
5		
6		7
7		

At the end there is a leaf of entries attached to the stem at 0, 1, and 7. Imagine that each leaf entry (single integer) occupies one unit of area, and hence count frequencies are proportional to areas, exactly as in the conventional histogram. Note that greater detail concerning number identify is preserved by the stem-leaf than is managed by the conventional histogram.

Stem-leaf displays are usually constructed by making a convenient but informal choice of leaf interval (here it was 10, but for more detail it could be 5, and for more smoothness, hence less detail, 20). A more formal approach, see Scott (1979), is that of picking the bin size,  $h_n$ , so as to minimize an integrated mean-squared error of estimate of the true density by the histogram. In summary, if the data is approximately Gaussian then the prescription turns out to be

$$h_n = \frac{3.49 s}{\sqrt{n}}$$

where  $h_n$  is bin size for a batch of  $n$ , and where  $s$  is the sample standard deviation. Observe that a few extreme outliers in an otherwise Gaussian-appearing batch will unjustifiably expand  $s$ , and over-coarsen the histogram. A robust estimate of  $\sigma$ , such as

$$\hat{\sigma} = \frac{(\text{Upper Quartile}) - (\text{Lower Quartile})}{1.35}$$

might then be recommended instead of  $s$ .

A second, more basic, feature of the raw data and the display is the crowding in the first class interval: numbers like 0\*\* are all jammed together, and there is evidence of systematic (right) skewness. In such cases transformation of the basic numbers before plotting is often useful, and in this particular case a first attempt might be with logarithms (base  $e = 2.7182\dots$ , but the particular base doesn't matter). Here are the logged numbers:

-0.083, 0.13, 0.69, 0.98, 1.87, 1.60, 1.02, 0.76, 0.13, 2.27  
 2.42, -1.39, 1.20, 2.54, -0.78, 1.02, -1.05, 4.20, -0.53, 0.61

Here we might try a scale of 0.5, so, rounding to the nearest tenth of an integer:

<u>Ranges</u>		<u>Stem-Leaf</u>	
-1.5	-2	-1	
-1.0	-1.5	-1	41
-0.5	-1.0	-0	85
-0.	-0.5	-0	1
0.	0.5	0	11
0.5	1.0	0	786
1.0	1.5	1	0020
1.5	2.0	1	96
2.0	2.5	2	34
2.5	3.0	2	5
3.0	3.5	3	
3.5	4.0	3	
4.0	4.5	4	2

Notice that the previous crowding and skewness has nearly disappeared, and we see revealed the vestigial appearance of two separate "humps," as well as the originally apparent outlier ( $\ln 66.67 \approx 4.2$ ). We are led to investigate the possibility that the data derive from two separate sources, with one exotic (mis?) measurement thrown in for good luck. In this form the data urges more upon us than it did originally.

## 2.2. Number Summaries

Traditionally it has been customary to summarize certain features of batches of measurements by moments: the (arithmetic) mean gives the "location" of the data set, the standard deviation summarizes "spread" (or "scale," or "width"), the third central moment measures "skewness," and so on. However, certain apparently more primitive measures are useful and have virtues.

Work with the observations after ordering them into

$$x_{(1)} < x_{(2)} < x_{(3)} < \dots < x_{(n)} .$$

a) Median,  $M$ .

Compute the median index  $m = (1+n)/2$ . If  $n$  is odd

$x_{(m)}$  is the middle (single) number, and if  $n$  even then average the two middle numbers. Thus the median is

$$M = \begin{cases} x_{(m)} & \text{if } n \text{ is odd ,} \\ \frac{1}{2}(x_{([m])} + x_{([m]+1)}) & \text{if } n \text{ is even .} \end{cases}$$

where  $[m]$  is the integer part of  $m$ . Notice that quite radical changes in extremes, e.g.  $x_{(1)}$  or  $x_{(n)}$ , effects  $M$  not at all; in fact changes from  $x_{(n)}$  to  $1000 x_{(n)}$ , creating a very isolated single observation, leaves  $M$  alone, so  $M$  is resistant (un-influenced) by such changes, or occurrence of outliers. On the other hand, the familiar mean responds dramatically and undesirably and is far from resistant. Thus the median is a useful candidate for "placing" or "locating" a rather concentrated batch of points when several exotic, separated, extremes are present; the mean is apt to strike a meaningless compromise. Of course, the median does not well-summarize a batch having two or more distinct but nearly equal-sized humps. But certainly it does no worse than the mean. Later we discuss some better measures, and also some for spread (alternatives to standard deviation).

(b) Quartiles: Lower =  $\underline{Q}$ , Upper =  $\bar{Q}$

Roughly one-quarter of the observations fall below  $\underline{Q}$  (above  $\bar{Q}$ ). Define

$$q = \frac{1}{2}(1 + [m]) .$$

Then

$$\underline{Q} = \begin{cases} x_{(q)} & \text{if } [m] \text{ is odd, as } q \text{ is integer ,} \\ \frac{1}{2}(x_{([q])} + x_{([q]+1)}) & \text{if } [m] \text{ is even .} \end{cases}$$

$\bar{Q}$  is defined analogously in terms of  $n - q + 1$ . In other words  $\underline{Q}(\bar{Q})$  is the median of the lower (upper) half of the ordered batch.

(c) Eighths: Lower =  $\underline{E}$ , Upper =  $\bar{E}$

About one-eighth of the observations fall below  $\underline{E}$  (above  $\bar{E}$ ). Let

$$e = \frac{1}{2}(1 + [q]) .$$

Then

$$\underline{E} = \begin{cases} x_{(e)} & \text{if } [q] \text{ is odd ,} \\ \frac{1}{2}(x_{([e])} + x_{([e]+1)}) & \text{if } [q] \text{ is even ,} \end{cases}$$

and  $\bar{E}$  is defined analogously in terms of  $n - e + 1$ .

(d) Extremes

There are simply

$$\underline{\text{Ext}} = x_{(1)} ,$$

$$\overline{\text{Ext}} = x_{(n)} .$$

A seven-number summary of the data is as follows:

M			
$\underline{Q}$	$\left( \frac{\underline{Q} + \bar{Q}}{2} \right) \equiv MQ$	$\bar{Q}$	$s_Q = \frac{MQ - M}{\bar{Q} - \underline{Q}}$
$\underline{E}$	$\left( \frac{\underline{E} + \bar{E}}{2} \right) \equiv ME$	$\bar{E}$	$s_E = \frac{ME - M}{\bar{E} - \underline{E}}$
$\underline{\text{Ext}}$	$\left( \frac{\underline{\text{Ext}} + \overline{\text{Ext}}}{2} \right) \equiv M\text{Ext}$	$\overline{\text{Ext}}$	$s_{\text{Ext}} = \frac{M\text{Ext} - M}{\overline{\text{Ext}} - \underline{\text{Ext}}}$

The quartile (eighth, extreme) means  $MQ(ME, M\text{Ext})$  can be quickly compared to  $M$  to detect systematic asymmetry or skewness.

Dimensionless measures of skewness are also given by the quantities  $s_Q, s_E, s_{\text{Ext}}$ .

Example:  $x_{(1)} = 1, x_{(2)} = 3, x_{(3)} = 5, x_{(4)} = 7, x_{(5)} = 9,$   
 $x_{(6)} = 111.$

Then  $m = \frac{1}{2}(1+6) = 3.5,$  so

$$M = \frac{1}{2}(5+7) = 6 .$$

Note that the mean  $\bar{x} = 22.67$ . This neither-fish-nor-fowl number has responded heavily to the isolated value 111. Next

$$q = \frac{1}{2}(1 + [3.5]) = 2 ,$$

so

$$\underline{Q} = 2, \quad \text{and} \quad \bar{Q} = 9.$$

Now

$$e = \frac{1}{2}(1 + [2]) = 1.5 ,$$

so

$$\underline{E} = \frac{1}{2}(1+3) = 2, \quad \text{and} \quad \bar{E} = \frac{1}{2}(111+9) = 60 .$$

The number summary then is

	6		
3	6	9	$s_Q = 0$
2	31	60	$s_E = \frac{30 - 6}{60 - 2} = 0.43$
1	56	111	$s_{Ext} = \frac{56 - 6}{111 - 1} = 0.45$

The numbers suggest positive skewness, but closer examination reveals that this is caused by the influence of one point alone.

Note that if the data is a sample from the Normal distribution with standard deviation  $\sigma$ , then convenient approximate estimates of  $\sigma$  are

$$(\bar{Q} - \underline{Q}) \frac{1}{1.35} = \hat{\sigma}_Q$$

$$\text{and} \quad (\bar{E} - \underline{E}) \frac{1}{2.30} = \hat{\sigma}_E$$

$$\text{and thus } \frac{\bar{E} - \underline{E}}{\bar{Q} - \underline{Q}} \approx 1.70$$

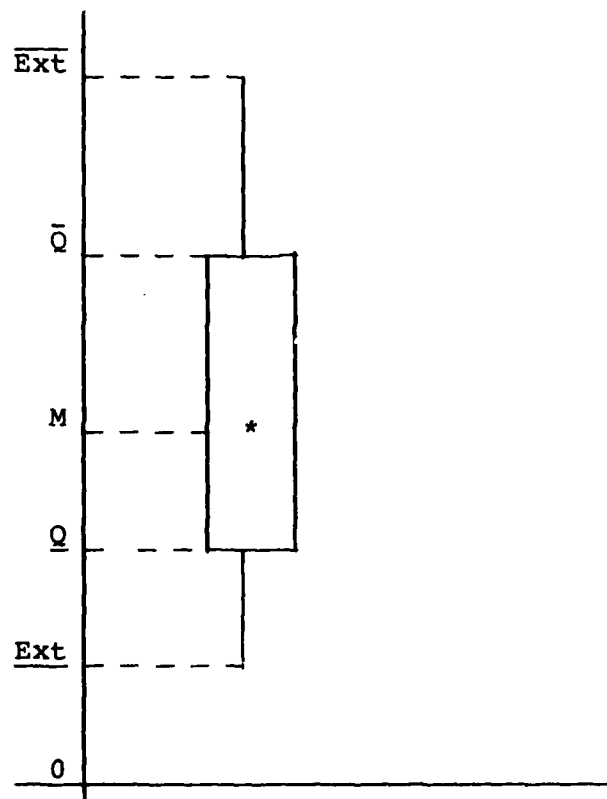
if data are approximately Normal; this is a handy check. An up-to-date test for precise Gaussian behavior is that of Wilk-Shapiro (1965). It involves a suitable linear combination of ordered observations as a test statistic. A cautionary note: critical values of the test statistic (rejection region) are obtained on the basis of independence of batch data values, often not an acceptable assumption in practice.

### 2.3. Box Plots

The box plot is a picture of the five-number (omit Eighths) summary. Simply draw a rectangular box with ends at  $\underline{Q}$  and  $\bar{Q}$ , and in which  $M$  is marked. In addition, connect up the extremes. As was done with stem-leaf we arrange the "box" vertically. Embellishments are sometimes useful: (i) it has been recommended that box width be proportional to  $\sqrt{n}$  to indicate effect of sample size when comparing different batches, (ii) warning points at  $\bar{Q} + 1.45(\bar{Q} - \underline{Q})$  and  $\underline{Q} - 1.45(\bar{Q} - \underline{Q})$ , something like "fences" of Tukey: if the data are Gaussian there is only about one percent of the data "outside" these values. Box plots are especially good for giving a quick appreciation of comparative distributional behavior for different batches.



# Simple Box Plot



#### 2.4. Rooted Histograms ("Rootograms")

Hark back to the histogram over bins of width  $h$ , with  $n_j$  observations in the  $j^{\text{th}}$  bin. The standard error of  $(n_j/n)$ , the frequency estimate of

$$\begin{aligned} p_j &= \text{Probability of an observation in the } j^{\text{th}} \text{ bin} \\ &= \int_{x_j}^{x_{j+1}} f(x) dx \approx f(x_j + h/2) h \end{aligned}$$

is at least under random sampling,

$$\{p_j(1-p_j)/n\}^{1/2} \approx \{(n_j/n)(1-n_j/n)n^{-1}\}^{1/2}$$

which suggests that the magnitude of the sampling fluctuations in the high-count (large  $p_j$ ) bins may be much larger than those in the low-count bins; on the other hand, the relative fluctuation magnitudes are greater in the low-count bins. The square-root transformation,  $\sqrt{n_j/n}$ , tends to stabilize (equalize) this variation:  $\sqrt{n_j/n}$  tends to estimate  $\sqrt{p_j}$  with standard error of about 0.5. This suggests several possibilities for useful graphical analysis:

- (1) Smoothing a raw histogram by smoothing  $\sqrt{n_j/n}$ -values and re-squaring the results, as possibly in

$$\left(\sqrt{\frac{n_j}{n}}\right)_S = \frac{1}{3} \left[ \sqrt{\frac{n_{j-1}}{n}} + \sqrt{\frac{n_j}{n}} + \sqrt{\frac{n_{j+1}}{n}} \right],$$

so the estimate of  $p_j$  becomes

$$\hat{p}_{j,S} = \left[ \left(\sqrt{\frac{n_j}{n}}\right)_S \right]^2.$$

Other approaches, such as running medians, (see Tukey (1977), p. 543 ff.) are likely to be effective. The attempt is to reduce sampling fluctuations without resorting immediately to the "ultimate smooth:" a simple model.

- (2) Assessing model fit by a hanging rootogram. Having a model distribution in mind, perhaps suggested by theory, we wish to graphically present the evidence for and against it using a histogram of data. To this end, plot

$$\Delta_j \equiv \sqrt{p_j} - \sqrt{\frac{n_j}{n}} \quad \text{vs} \quad x_j + \frac{h}{2} ;$$

this is equivalent to looking at a plot of the square root of the density, with roots of the histogram values hanging from it. If the model is in basic agreement with the data then about 95% of the values of the differences  $\Delta_j$  should lie within unit distance of zero. These differences "should" also be nearly pattern-free as order goes. Departures vividly show exactly where the model and the data disagree. Thus the procedure has a focus, and a specific diagnostic slant. It can clearly be used for histograms in more than one dimension.

The above idea, also due to Tukey can be extended to supply a formal test of goodness of fit analogous to the classical Chi-squared test. At the moment the concern is with graphically exposing data features and with indications of model-data discrepancies.

### 3. More Summary Measures

In Section 2 we worked with the median,  $M$ , as a summary measure of location, and with the midspread  $\bar{Q} - Q$  as a summary measure of distributional spread. Of course the classical measures would be the mean,  $\bar{x}$ , and standard deviation,  $s (= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2})$ . The latter are suspect because of their sensitivity to occurrence of a few wild values--perhaps recording errors. Here are some currently well-regarded alternatives.

#### 3.1. Winsorized mean; confidence limits

Suppose a set of nearly symmetrically distributed observations are in hand. If a small number of outlying observations are feared one can diminish their impact by symmetrically Winsorizing to level  $g$  ( $g \leq 0.20n$ ,  $n$  being number in the batch):

(a) Order the batch

$$x_{(1)} < x_{(2)} < \dots < x_{(g)} < x_{(g+1)} < \dots < x_{(n-g)} < x_{(n-g+1)} < \dots < x_{(n)}.$$

(b) Winsorize, level  $g$ :

$$y_{(1)} = y_{(2)} = \dots = y_{(g)} = x_{(g+1)} < y_{(g+1)} = x_{(g+1)} < y_{(g+2)} < \dots$$

$$y_{(n-g)} = x_{(n-g)} = y_{(n-g+1)} = y_{(n-g+2)} = \dots = y_{(n)}.$$

In other words, define the  $g$  smallest in the Winsorized batch ( $y$ 's) to equal the  $g^{\text{th}}$  smallest raw data point,  $x_{(g)}$ . Do the same symmetrically at the upper end. The result will be tied values ( $g$  at bottom,  $g$  at top).

(c) Average, to get a g-Winsorized mean:

$$\bar{x}_{Wg} = \frac{1}{n} \sum_{i=1}^n y_i$$

Notice that if  $n$  is odd and you Winsorize to the extent  $g = \frac{n-1}{2}$  the result is precisely the median. The Winsorized mean can be viewed as a broadened median, with a median's virtues (oblivious to outliers), but that uses more of the sample information. Winsorizing is named for C. P. Winsor, an insightful applied statistician.

(d) Confidence limits using the Winsorized mean.

Confidence limits attempt to express the uncertainty inherent in a particular estimate of a fixed quantity (such as mean radiation, mean visibility, etc.). If one is estimating the mean,  $\mu$ , of a symmetric distribution (possibly after transformation), and if the observations are independent, then it is appropriate to use the classical Student's  $t$  (or a Gaussian approximation): with confidence (not probability) of  $(1-\alpha) \cdot 100\%$ ,

$$\bar{x} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}}$$

where  $t_{\beta}(n)$  is the  $\beta \cdot 100\%$  point of a Student's  $t$  with  $n$  "degrees of freedom;"  $\bar{x}$  and  $s$  are the sample mean and standard deviation, respectively. Such intervals are reasonably satisfactory even if the data are not quite Gaussian, but if there are some extreme outliers,  $s$  blows up and the intervals are much

too wide. One should also compute the Winsorized limits: after computing  $\bar{x}_{wg}$ , find

$$s_{wg}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_{(i)} - \bar{x}_{wg})^2$$

and then find the confidence limits

$$\bar{x}_{wg} + t_{\alpha/2, (n-2g-1)} \frac{s_{wg}}{\sqrt{n}} \leq \mu \leq \bar{x}_{wg} + \left( \frac{n-1}{n-2g-1} \right) t_{1-\alpha/2, (n-2g-1)} \frac{s_{wg}}{\sqrt{n}} .$$

These may be decisively narrower; if so, look for outliers. Pick a succession of  $g$ -values, but not to exceed  $0.20n$ .

For more details, see Dixon and Tukey (1968) and Dixon and Yuen (1973).

### 3.2. Biweights

The Winsorizing procedure has to a degree been supplanted by a procedure called biweights. This involves an iterative least-squares calculation with weights on individual observations that decrease as the degree apparent non-representativeness (exoticizm) of a data point increases. Let  $x_1, x_2, \dots, x_n$  be data points; then here is a recipe for fitting a resistant center via biweights: referring to the  $k+1^{\text{st}}$  iteration in terms of the  $k^{\text{th}}$ ,

$$\bar{x}(k+1) = \frac{\sum_{i=1}^n w_i(k) \cdot x_i}{\sum_{i=1}^n w_i(k)}$$

where the weight of observation  $i$  is

$$w_i(k) = \begin{cases} \left[ 1 - \left( \frac{x_i - \bar{x}(k)}{cs_k} \right)^2 \right]^2, & \text{if } \left( \frac{x_i - \bar{x}(k)}{cs_k} \right)^2 < 1 \\ 0 & \text{otherwise.} \end{cases}$$

and  $s_k$  is a spread-measure, perhaps

$$s_k = \text{median}\{|x_i - \bar{x}(k)|\}$$

the so-called MAD or median absolute deviation, or alternatively

$$s_k = \bar{Q}_k - \underline{Q}_k$$

the midspread of the residuals at the  $k^{\text{th}}$  stage. These estimates are essentially equivalent. Here  $c$  is a tuning parameter; a value of  $c = 6$  or  $9$  has been recommended for general use, while  $c = \infty$  yields up the OLS estimates.

#### 4. Relations Between Variables, Models, and Model Assessment

The discovery and establishment of quantitative relationships between measurable, or classifiable, variables is a primary scientific concern. Once a relationship is uncovered it is soon likely to be thought to be potentially useful for some human purpose, and attempts will be made to apply it. Many examples exist in the remote sensing area. Of course the quality, or strength, of the relationship must be such as to make its application feasible. Statistical methods and thinking are frequently used, sometimes very informally, to help in finding a relationship and to expressing it in succinct, often, formally mathematical terms (as a mathematical model), to characterizing its deficiencies or biases, and to expressing the uncertainties inherent in its use. Here is a brief review of some of these methods.

##### 4.1. Graphical Plotting for Relationship Exploration

Many relationships between variables are initially guessed from scientific theory; or at least from subject-matter knowledge, intuition, or low cunning. When relevant data becomes available the obvious first step is to plot it graphically, if possible. This is easy if the relationship of interest is between two variables:  $x$ , an explanatory variable, or condition (sometimes called factor, or lately carrier) and  $y = f(x)$ , a response. Suppose one has observed pairs of these variable values:  $y_i, x_i; i=1,2,\dots,n$ . Then plots of  $f(x_i)$  vs  $x_i$  and  $y_i$  vs  $x_i$  on the same graph (e.g. rectangular coordinates) are sometimes presented for comparison. Since the range of  $f(x)$ --and  $y$ --may be considerable, such plots often obscure real systematic differences, it is often wise to plot and study residuals



$$r_i = y_i - f(x_i)$$

either vs  $x_i$ , or vs  $y_i = f(x_i)$ ; plotting residuals vs estimated values  $\hat{y}_i$  is one of the obvious ways of evaluating a relationship depending on several explanatory variables. Examination of a plot often reveals some systematic departure from linearity, such as a definite curvature may become evident; frequently this may be essentially removed by transformation (alternatively called re-expression). This is a good idea, for comparison of plots to straight lines (looking at residuals from straight lines) seems well adapted to human perceptions. The powers  $y \rightarrow y^p$  ( $p > 1$ , or  $p < 1$ ) or  $y \rightarrow \ln y$  (zero power) are often recommended for transformation to the nearly linear. Of course theory (e.g. the laws of physics, such involving squares, etc.) may point the way to an approximate linear plot. Then systematic deviations may be identified and interpreted on their own merits.

#### 4.2. Relationship Fitting

Although the mathematical form of the relationship between response  $y$  and explanatory variable  $x$  may suggest itself from theory, certain constants (parameters) nearly always require numerical determination. This means that a mathematical model exists, and must be fit to the data. Afterwards, one can look at residuals for indications of mis-fit, or apply the model to make predictions and check out its errors. One can sometimes segment the data, utilize a part to fit the model, and then use the remainder to examine the performance of the fitted model. This latter procedure is called cross-validation. Again residual plots are desirable diagnostics.

The actual fitting (constant, or parameter, estimation) problem can be carried out in various ways. We illustrate in terms of a postulated simple linear relation between response,  $y$ , and explanatory variable,  $x$ , i.e.

$$y = a + bx ,$$

with  $a$  and  $b$  to be determined. Here are some options.

- o Simple least squares
- o Median, two-points
- o Biweights

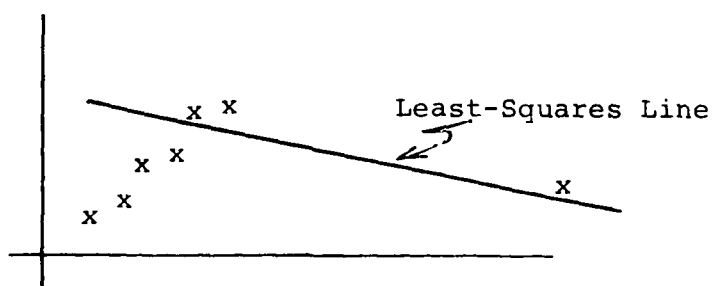
Given a set of  $(y_i, x_i)$  data, least squares proceeds by minimizing the sum of squares

$$\sum_{i=1}^n (y_i - a - bx_i)^2$$

by choice of  $a$  and  $b$ . Linear equations result for estimated  $a$ , namely  $\hat{a}_{LS}$  and estimated  $b$ , namely  $\hat{b}_{LS}$ . Many computer programs exist for doing this problem, and also for doing the multiple regression problem: fit

$$y_j, j=1, 2, \dots, n \text{ by } \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_p x_{pj}.$$

For better or worse, any set of data can be conveniently fitted by such a linear function, i.e. the parameters  $\beta_0, \beta_1, \dots, \beta_p$  can essentially always be determined numerically, whether or not the postulated relationship makes sense. Furthermore, the estimates,  $\hat{\beta}_i$ , turn out to be linear functions of the  $y_j$  values. This suggests that  $\hat{\beta}_i$  values also respond linearly to changes--perhaps even unfortunate or unsuspected discrepant values or errors--in the  $y_j$  values, and so a few funny values can give a warped picture:



The fitted line may not follow the main data cluster, but instead fasten itself to a single exotic point. For some discussion of this and remedies see Mosteller and Tukey (1977) and also Belsley, Kuh, and Welsch (1980).

The second method mentioned ("median, two-points") avoids some of the above difficulties. Prescription: order the explanatory variables from smallest to largest:  $x_{(1)} < x_{(2)} \dots$

$x_{(n/3)} < x_{(n/3+1)} < \dots x_{(2n/3)} < \dots x_{(n)}$ . Call the set of values  $x_{(1)}, x_{(2)}, \dots, x_{(n/3)}$  the low group, and  $x_{(2n/3)}, x_{(2n/3+1)}, \dots, x_{(n)}$ , the high group. Let  $x_\ell$  be the median of the low group, and  $x_h$  be the median of the high group. To each member of the low group there is a  $y$ ; let  $y$  be the median of the corresponding low  $y$ 's, and  $y_h$  be the median of the high  $y$ 's. Now fit:

$$\hat{b}_{MED} = \frac{y_h - y_\ell}{x_h - x_\ell}.$$

$$\hat{a}_{MED} = y_h - \hat{b}_{MED} x_h$$

This procedure is a modification of an ancient procedure that utilizes means instead of medians. Since medians are more resistant to outliers than means, this procedure is a move in the right direction.

The third option ("biweights") is an iteratively weighted least squares approach. It has been programmed for many computers, including even a hand-held TI-59. This procedure develops a weight for each observation; the weight diminishes as the observation becomes apparently discrepant, as in the earlier discussion. Details are omitted here.

Example. Sea-surface wind speed and white-cap coverage.

It has been proposed by Monahan (1971) that magnitudes of sea-surface winds and white caps are strongly related. Presumably white-cap coverage can be assessed remotely, and wind speeds deduced there from. A set of data offered by Toba and Chaen (1971) has been analysed using the above methods. An

appealing functional form is the power law

$$y = \alpha x^{\beta} ;$$

it has been argued on theoretical grounds;  $\beta$  is supposed to be in the neighborhood of 3. A computer scatter plot shows that this is plausible. An initial cube-root transformation ( $y^{1/3}$  vs  $x$ , equivalent to the model  $y = \alpha x^3$ ) seems to straighten the plot reasonably well, but note that a few zero values occur; other outliers are less obvious.

It is natural to try to fit the log-transformed version of the power law, for this is now linear:

$$\ln y = \ln \alpha + \beta \ln x$$

Here are some results (OLS means Ordinary Least Square,  $B_i(i)$  means the biweight result after  $i$  iterations,  $s_i$  means robust scale after  $i$  iterations); note that it has been necessary to fit  $\ln(y + \text{start})$ ,  $\text{start} = 0.001$  here, in order to avoid the embarrassment of logging exact zeros. The "start" value can be chosen wisely; no attempt to do so is exhibited here, however.

#### PARAMETER FITS FOR LOG-LINEAR MODEL OF WHITE CAPS vs WIND

Method	Estimate		
	$\ln \alpha$	$\beta$	$s_i$
OLS = $B_i(1)$	-9.65	3.51	1.44
OLS (zeros out)	-10.31	4.22	----
$B_i(2)$	-10.73	4.24	1.26
$B_i(4)$	-11.32	4.68	1.05
$B_i(6)$	-11.39	4.25	1.03

Apparently the biweight calculation yields quite different results from OLS, even after visual culling of apparent zeros has been conducted. It would be of interest to utilize the techniques of Cook (1977) and Belsley, Kuh, and Welsch (1980) to further identify influential observations. Of course the present setup is simple enough so that visual examination will reveal most of what is present. This would not necessarily be so if the explanatory variable were a vector.

Some computer plots are included to show that the biweight procedure provides fits that tend to dramatically reveal the presence of outliers. In many cases the presence of such outliers represents opportunity for new insight and discovery, as is pointed out by Tukey (1977). Outliers should never be immediately "thrown away," but rather are candidates for special attention. See Kruskal (1960) for some wise discussion of this issue.

#### 4.3. Fitting Probabilities and the Like

It is often appealing to explore the relationship between the probability of some environmental (or other) event and some reasonable explanatory variables. This amounts to estimating conditional probabilities. Often the events in question are of the form "rain," "fog," "visibility in the range 0-5 kilometers," etc. Candidate explanatory variables may be the product of a remote sensing system, or a numerical weather forecasting scheme ("model output statistics"), or a combination thereof. Some attempts to use persistence ("it rained in Monterey on January 14, so the probability of rain on January 15 is  $\approx 0.2$ , whereas without rain it is  $\approx 0$ ") are also appealing. In the latter form the simple idea of Markov chains has been employed.

Ordinary linear regression has been used to predict event probabilities, e.g. the REEP scheme of R. G. Miller (1964). This has the aesthetic difficulty that probabilities are between zero and 1, which ordinary regression doesn't recognize. Two ways of dealing with probability regression problems are as follows:

- o logistic regression
- o conditional probabilities.

In logistic regression one utilizes the simple model

$$P(\text{event } E | \text{explanatory variable } X=x) = p(x) = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

where  $a$  and  $b$  are constants to be determined. Although this model has been written in terms of one explanatory variable, multiple regression options are available. One can contemplate fitting the logistic model in several ways:

(i) By maximum likelihood. Suppose the explanatory variable takes on value  $x_i$  on the  $i^{\text{th}}$  occasion, and a success (event, e.g. rain) occurs,  $i = 1, 2, \dots, n$  happenings occur. Then the likelihood of the  $a, b$  using the data  $(\delta_i, x_i, i=1, 2, \dots, n)$ , where  $\delta_i = \begin{cases} 1 & \text{if event occurs,} \\ 0 & \text{otherwise} \end{cases}$  is seen to be

$$L(a, b; \underline{x}) = \prod_{i=1}^n \left( \frac{e^{a+bx_i}}{1 + e^{a+bx_i}} \right)^{\delta_i} \left( \frac{1}{1 + e^{a+bx_i}} \right)^{1-\delta_i}$$

or

$$\ell(a, b; \bar{x}) = \ln L = \sum_{i=1}^n \delta_i (a+bx_i) - \sum_{i=1}^n \ln[1 + e^{a+bx_i}]$$

Differentiation, then solution for  $a, b$  yields maximum likelihood parameter estimates. See Cox (1970), and, recently, McCullagh (1980). Pregibon (1980) has generated robust procedures.

(ii) By grouping. If the data set is large one can order the  $x$ 's, split the ordered values into an equal number of groups,  $g$ , calculate the frequency of successes in each group  $\hat{p}_j = \frac{n_j}{(n/g)}$  where  $n_j$  is the number of successes in group  $j$  and  $n$  is the total number of observations. Now let  $x_j$  be a representative explanatory variable value for group  $j$ ; the median may be appropriate. Note that

$$y_j = \ln(\hat{p}_j / (1 - \hat{p}_j)) \quad \underline{\text{vs}} \quad a + b x_j$$

should be nearly linear if the logistic relationship holds (if not, try a transformation). In any case it is linear in the parameters, so a fit can be readily made and diagnostically viewed. One can even fit the above by weighted least squares, or perhaps biweights. Such procedures are under investigation at the Naval Postgraduate School by the author and Dennis Mar; they seem promising. See Cox (1970), especially Chap. 3, for some good discussion. The procedure can be made multivariate ( $x$  a vector), provided there is a good deal of data.

Notice that the above method handles only dichotomous situations ("rain, no rain," "visibility in category (\*), visibility in category (not (\*))"). A multiclass version of the above has been devised by P. Bloomfield, J. Lehoczky and the author (possibly also by others) and is under development.



Another approach to the multiclass problem is by conditional probabilities. Suppose  $E_j$  is the event that an observation (e.g. of visibility state) lies in class  $j$  ( $j = 1, 2, \dots, C$ ), and  $x_j$  is the corresponding explanatory variable value. One can construct an estimate of the density function given that  $E_j$  has occurred,  $f(x_j|E_j)$ , for each class,  $j = 1, 2, \dots, C$ ; the histogram, raw or smoothed (c.f. Wegman (1979)), or a simple model (normal, lognormal, etc.) may do well. By Bayes' theorem, then

$$P\{E_j|x\} = \frac{f(x|E_j)P\{E_j\}}{\sum_{j=1}^C f(x|E_j)P\{E_j\}} ;$$

here  $P\{E_j\}$  is the overall marginal probability of an event in class  $j$ , estimable from history. Such an approach is under empirical investigation by the author, utilizing some meteorological data.

## 5. Time Series

Many of the data sets encountered in studies of the natural environment have distinctive time-series structure. This means that successive observations in time (and contiguous observations in space) are likely to be rather similar or correlated as a result of both (a) important natural phenomena associate systematically with seasons, years, geographical places, and forces of nature, and also (b) what may be considered to be the haphazard superposition of a variety of additional effects, among which may be measurement error. These latter effects typically do not appear independent from observation to observation, a fact that represents both opportunity and difficulty in statistical analysis. In the present section we review a few basic notions and concepts in time series analysis. The topic is subtle, and deserves more than it gets in this discussion.

### 5.1. Components of a Time Series

One way of thinking about a time series, say of monthly total precipitation at a point in space, or visibility at a point, is in terms of the simple components

- |   |                                |                     |
|---|--------------------------------|---------------------|
| o | (apparent) trend               | } systematic effect |
| o | (apparent) seasonal effect     |                     |
| o | (random) disturbance or noise. |                     |

The trend ideally represents a systematic long-term change; "apparent" is appended because a steady trend may actually be quite impermanent, giving way to a new and different trend. Think of the economy, particularly the stock market indices. Mean sea levels measured by tide gauges are also of this nature, possibly

because a regime of slow, regular, changes in land supporting a tide gauge may rather suddenly be supplanted by a different regime, perhaps the result of human activity. Sometimes a trend may be taken to be linear, at least provisionally. The seasonal effect often seems inevitable: ordinarily it rains in the winter in Monterey, and not in the summer. Tides behave in accordance with time of day, with a slow trend superimposed. Left over is the contribution of random disturbance or noise, which represents the joint influence of other conditions, including measurement error. Interestingly, the latter random component has received more sophisticated mathematical-statistical attention than have the other components. Understanding trends and seasonals (apparently systematic effects) probably depends upon understanding the underlying subject-matter area.

## 5.2. Decomposition

Investigation of a time series  $z_t$ ,  $t = 1, 2, \dots$  ought to begin by a study of graphical display. The inter-relations between several series is often suggested by such an approach. Suppose one wishes to isolate the semi-permanent component of the series (trend and seasonal). This can be approached in the following ways:

- o Fit a specific function to the trend,
- o Smooth, either non-robustly or robustly.

If a plot shows a nearly linear trend (perhaps after transformation, e.g. by logs) one can fit

$$z_t \text{ vs } a + bt \qquad t = 1, 2, \dots, T$$

by least squares. Now if the true situation were well-represented by the model

$$z_t = a + bt + \epsilon_t$$

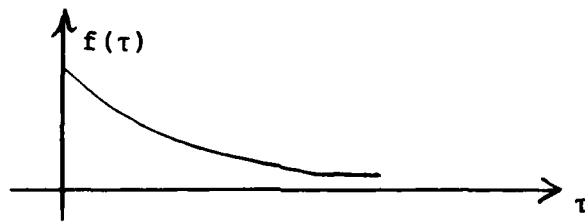
with  $\{\epsilon_t, t=1,2,\dots\}$  a sequence of non-independent random variables with, nevertheless  $E[\epsilon_t] = 0$ ,  $\text{var}[\epsilon_t] = \sigma^2$ , the Ordinary Least Squares is less than perfectly efficient. However it is ordinarily a useful approach--maybe all that is available. Once  $\hat{a}_{LS}$ ,  $\hat{b}_{LS}$  are available one can subtract away the fit, leaving the residuals

$$r_t \equiv z_t - \hat{a}_{LS} - \hat{b}_{LS}t \ (\approx \epsilon_t).$$

for study. If the trend removal has been effective, and no intervening seasonal has occurred, the  $r_t$ 's will have roughly constant variation for all  $t$  (box plot segments of the series for diagnosis), and a characterization or study of the  $r_t$ 's is in order. Note that if the covariance function

$$f(\tau) = \frac{1}{T-\tau} \sum_{t=1}^{T-\tau} r_t r_{t+\tau}$$

remains and diminishes only slowly as  $\tau$  increases this signifies substantial similarities between observations separated in time contributed by something (here it is assumed to be the noise component, although it might well be an un-removed contribution by seasonal, or a change in trend). Often  $f(\tau)$  looks something like this



### 5.3. Noise Models

It is convenient to employ simple parametric models to describe noise behavior. Here are several

- o Simple Markov or Autoregressive:  $\epsilon_t = \rho \epsilon_{t-1} + a_t$ ,  
 $\{a_t, t=1,2,\dots\}$  is "white noise," a sequence of independently Gaussian variables.
- o Simple Moving Average:  $\epsilon_t = \frac{a_t + a_{t-1} + \dots + a_{t-p}}{p+1}$   
 $p$  an integer  $\geq 0$ .
- o Autoregressive-Moving Average (ARMA (1,1)):

$$\epsilon_t - \rho \epsilon_{t-1} = a_t - \theta a_{t-1}$$

It is well to start with the simple Markov, for which one may estimate  $\rho$  from the residuals by

$$\hat{\rho} = \frac{1}{T-1} \sum_{t=1}^T r_t r_{t-1}$$

The theoretical autocovariance function of the Markov (AR) is actually

$$f_\tau = \sigma^2 \rho^{|\tau|} \quad |\tau| = 0, 1, 2, \dots$$

so as a check  $(\hat{\rho})^\tau$  should die off exponentially. A plot on semi-log paper is helpful for a check.

Although least squares may provide reasonable estimates for parameters  $a$  and  $b$ , the attempt to use the sum of the squares of the residuals to estimate  $\sigma^2$  (noise variance), and then to use this estimate in conventional formulas for standard errors and confidence limits for  $a$  and  $b$  is often a bad mistake: the "standard" results, such as those found in regression packages at computer centers, are likely to give wildly optimistic (falsely precise) results. Such effects have been noticed by Abreu (1980) in an investigation of sea level trends on the U.S. Pacific Coast.

Note that if a linear trend has been fitted by OLS, and the residuals appear to be  $AR(1)$ , then the variance of the slope term estimate,  $\hat{b}_{OLS}$ , is approximately that given by the OLS formulas appropriate for purely random noise, multiplied by  $\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$ . For more detail see Bloomfield (1980).

## APPENDIX

The purpose of this appendix is to report the results of simulation tests of the performance of Winsorized t confidence limits as compared to ordinary Student's t. In particular, comparative performance is reported, as measured by (i) confidence interval coverage (fraction of intervals actually containing the true value of the mean), and (ii) confidence interval average (estimate of expected) width.

o Simulated Data. The "data" were obtained as follows. Alternatively,

- (a) Samples of size  $n = 20$  from Normal  $(0,1)$ . The random variable of which the observations are independent instances will be called  $Z$ .
- (b) Samples of size  $n = 20$  from the distribution of  $Y = Ze^{hZ^2}$ ,  $h > 0$ . The  $y$ -values are more spread out (but still symmetrically so) than are the basic  $z$ 's, in order to represent long, fat, tails possibly resulting from outliers. The above convenient form has been suggested by Tukey.
- (c) Samples of size  $n = 20$  from the distribution of  $W = (e^{g'Z} - 1)/g'$ , the latter being recognizable as the asymmetric log-Normal form.

o Experimental Sampling. Suppose that a sample is available (from one of the above situations, but the exact source being unknown), compute confidence limits for the mean of the parent

distribution, using (A) ordinary Student's  $t$ , and (B) Winsorized  $t$ , using  $g = 2$ , i.e. the lowest two values were given the value of  $x_{(3)}$ , and the highest two values the value of  $x_{(18)}$ . Do so repeatedly ( $k = 1000$  times at present) and compare for (i) coverage to nominal (here 95% two-sided) and (ii) average width of confidence intervals. These parameter values were examined:

$$g' = 0.2, \quad h = 1.00, \quad \text{and} \quad h = 2.00$$

o Results of the Sampling Experiment.

<u>Case</u>	<u><math>g'</math></u>	<u><math>h</math></u>	<u>Method</u>	<u>Coverage(%)</u>	<u>Av. Width</u>
Normal (Z)	0	0	Student	95.5	0.92
Long-Tail (Y)	0	1.00	Student	99.3	8658
Skewed (W)	0.2	0	Student	94.8	0.95
Long-Tail (Y)	0	1.00	Winsor(2)	96.4	8.46
Skewed (W)	0.2	0	Winsor(2)	95.3	0.98
Long-Tail (Y)	0	2.00	Student	99.8	$4.3 \times 10^{10}$
Long-Tail (W)	0	2.00	Winsor(2)	98.0	548

o Conclusions. Obvious indication of the efficacy of Winsorizing long-tailed (Y) observations as compared to traditional student's  $t$ . At that, the level of Winsorization ( $g = 2$ ) is probably too small, and a larger value would provide valid narrower, intervals. Alternatively, utilize biweights.



# REFERENCES AND BIBLIOGRAPHY

- [1] Abreu, F. A. T. V. (1980). "Determination of land elevation changes using tidal data," Naval Postgraduate School thesis for the M.S. Degree in Oceanography. (W. Thompson and D. P. Gaver, Advisors).
- [2] Belsley, D., Kuh, E., Welsch, R. (1980). Regression Diagnostics. John Wiley & Sons, New York, NY.
- [3] Bloomfield, P. (1980). "Trend estimation with autocorrelated errors, with physical application." Three lectures delivered at the Naval Postgraduate School, Monterey, California.
- [4] Cook, R. (1977). "Influential observations in linear regression," Technometrics, pp. 15-18.
- [5] Cox, D. R. (1970). The Analysis of Binary Data. Chapman and Hall, London, England.
- [6] Dixon, W. J. and Tukey, J. W. (1978). "Approximate behavior of the distribution of Winsorized  $t$ ," Technometrics 10.
- [7] Kruskal, W. H. (1960). "Some remarks on wild observations," Technometrics, Vol. 2, No. 1, pp. 1-3.
- [8] Launer, R. L. and Wilkinson, G. N. (eds.) (1979). Robustness in Statistics.
- [9] McCullagh, P. (1980). "Regression models for ordinal data," J. of the Royal Statistical Society, B., pp. 109-142.
- [10] McGill, R. M., Tukey, J. W. and Larsen, W. A. (1978). "Variations of box plots," American Statistician, Vol. 32, pp. 12-16.
- [11] Miller, R. G. (1964). "Regression estimation of event probabilities," Technical Report 7411-121, The Travelers Research Center, Inc., Hartford, Connecticut.
- [12] Monohan, E. C. (1971). "Oceanic Whitecaps," J. of Physical Oceanography, 1, pp. 139-144.
- [13] Mosteller, F. and Tukey, J. W. (1977). Data Analysis and Regression, Addison-Wesley Publishing Co., Reading, Mass.
- [14] Pierce, D. A. (1980). "A survey of recent developments in seasonal adjustment," The American Statistician, 34, pp. 125-134.
- [15] Pregibon, D. (1980). "Resistant fits for some commonly used logistic models with medical applications," Technical Report, Statistics Department, Princeton University, Princeton, NJ.

- [16] Pregibon, D. (1980). "Logistic regression diagnostics," Research paper, available through Statistics Department, Princeton University, Princeton, NJ.
- [17] Scott, D. W. (1979). "On optimal and data-based histograms," Biometrika, Vol. 66, No. 3, pp. 605-610.
- [18] Toba, Y. and Chaen, M. (1973). "Quantitative expression of the breaking of wind waves on the sea surface," Records of Oceanographic Works in Japan, 12, No. 1, pp. 1-11.
- [19] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley Publishing Co., Reading, Mass.
- [20] Wegman, E. (1972). Nonparametric density estimation: I. A Summary of Available Methods. Technometrics, 14, pp. 533-546.
- [21] Wegman, E. (1979). Notes on Time Series.
- [22] Wilk, M. B. and Shapiro, S. S. (1965). "An analysis of variance test for normality (complete samples)," Biometrika, 52, pp. 591-611.
- [23] Yuen, K. K. and Dixon, W. J. (1973). "The approximate behaviour and performance of the two-sample trimmed  $t$ ." Biometrika, 60, pp. 369-374.

# INITIAL DISTRIBUTION LIST

	Copies		Copies
Statistics and Probability Program (Code 436) Office of Naval Research Arlington, VA 22217	3	Office of Naval Research Scientific Liaison Group Attn: Scientific Director American Embassy - Tokyo APO San Francisco 96503	1
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2	Applied Mathematics Laboratory David Taylor Naval Ship Research and Development Center Attn: Mr. G. H. Gleissner Bethesda, Maryland 20084	1
Office of Naval Research New York Area Office 715 Broadway - 5th Floor New York, New York 10003	1	Commandant of the Marine Corps (Code AX) Attn: Dr. A. L. Slafkosky Scientific Advisor Washington, DC 20380	1
Commanding Officer Office of Naval Research Eastern/ Central Regional Office Attn: Director for Science 666 Summer Street Boston, MA 02210	1	Director National Security Agency Attn: Mr. Stahly and Dr. Maar (R51) Fort Meade, MD 20755	2
Commanding Officer Office of Naval Research Western Regional Office Attn: Dr. Richard Lau 1030 East Green Street Pasadena, CA 91101	1	Navy Library National Space Technology Laboratory Attn: Navy Librarian Bay St. Louis, MS 39522	1
Commanding Officer Office of Naval Research Branch Office Attn: Director for Science 536 South Clark Street Chicago, Illinois 60605	1	U.S. Army Research Office P.O. Box 12211 Attn: Dr. J. Chandra Research Triangle Park, NC 27706	1

	Copies		Copies
OASD (I&L), Pentagon Attn: Mr. Charles S. Smith Washington, DC 20301	1	ATAA-SL, Library U.S. Army TRADOC Systems Analysis Activity Department of the Army White Sands Missile Range, NM 88002	1
ARI Field Unit-USAREUR Attn: Library c/o ODCSPER HQ USAEREUR & 7th Army APO New York 09403	1	Dr. Edward J. Wegman Statistics and Probability Program Office of Naval Research Arlington, VA 22217	1
Naval Underwater Systems Center Attn: Dr. Derrill J. Bordelon Code 21 Newport, Rhode Island 02840	1	Library (Code 0142) Naval Postgraduate School Monterey, CA 93940	2
Library, Code 1424 Naval Postgraduate School Monterey, CA 93940	1		
Technical Information Division Naval Research Laboratory Washington, DC 20375	1		
Dr. Barbara Bailar Associate Director, Statistical Standards Bureau of Census Washington, DC 20233	1		
Director AMSAA Attn: DRXSY-MP, H. Cohen Aberdeen Proving Ground, MD 21005	1		
Dr. Gerhard Heiche Naval Air Systems Command (NAIR 03) Jefferson Plaza No. 1 Arlington, VA 20360	1		
B. E. Clark RR #2, Box 647-B Graham, NC 27253	1		
Leon Slavin Naval Sea Systems Command (NSEA 05H) Crystal Mall #4, Rm. 129 Washington, DC 20036	1		

	Copies		Copies
Technical Library Naval Ordnance Station Indian Head, MD 20640	1	Mr. Jim Gates Code 9211 Fleet Material Support Office U.S. Navy Supply Center Mechanicsburg, PA 17055	1
Bureau of Naval Personnel Department of the Navy Technical Library Washington, DC 20370	1	Mr. Ted Tupper Code M-311C Military Sealift Command Department of the Navy Washington, DC 20390	1
Library Naval Ocean Systems Center San Diego, CA 92152	1	Mr. F. R. Del Priori Code 224 Operational Test and Evaluation Force (OPTEVFOR) Norfolk, VA 23511	1
Defense Logistics Studies Information Exchange Army Logistics Management Center Attn: Mr. J. Dowling Fort Lee, VA 23801	1	Professor D. P. Gaver Department of Operations Research Naval Postgraduate School Monterey, CA 93940	20
Reliability Analysis Center (RAC) RADC/RBRAC Attn: I. L. Krulac Data Coordinator/ Government Programs Griffiss AFB, New York 13441	1	Professor Barnard H. Bissinger Mathematical Sciences Capitol Campus Pennsylvania State University Middletown, PA 17057	1
Dr. M. J. Fischer Defense Communications Agency Defense Communications Engineering Center 1860 Wiehle Avenue Reston, VA 22090	1	Professor Robert Serfling Department of Mathematical Sciences The Johns Hopkins University Baltimore, Maryland 21218	1
Mr. David S. Siegel Code 260 Office of Naval Research Arlington, VA 22217	1	Professor Ralph A. Bradley Department of Statistics Florida State University Tallahassee, FL 32306	1

Copies

Copies

Professor G. S. Watson  
Department of Statistics  
Princeton University  
Princeton, NJ 08540

1

Professor H. Chernoff  
Department of Mathematics  
Massachusetts Institute of Technology  
Cambridge, MA 02139

1

Professor P. J. Bickel  
Department of Statistics  
University of California  
Berkeley, CA 94720

1

Professor D. O. Siegmund  
Department of Statistics  
Stanford University  
Stanford, CA 94305

1

Professor F. J. Anscombe  
Department of Statistics  
Yale University  
Box 2179 - Yale Station  
New Haven, CT 06520

1

Professor Grace Wahba  
Department of Statistics  
University of Wisconsin  
Madison, Wisconsin 53706

1

Professor S. S. Gupta  
Department of Statistics  
Purdue University  
West Lafayette, Indiana 47907

1

Professor Walter L. Smith  
Department of Statistics  
University of North Carolina  
Chapel Hill, NC 27514

1

Professor R. E. Bechhofer  
Department of Operations Research  
Cornell University  
Ithaca, New York 14850

1

Professor S. E. Fienberg  
Department of Statistics  
Carnegie-Mellon University  
Pittsburgh, PA 15213

1

Professor D. B. Owen  
Department of Statistics  
Southern Methodist University  
Dallas, Texas 75275

1

Professor Gerald L. Sievers  
Department of Mathematics  
Western Michigan University  
Kalamazoo, Michigan 49008

1

Professor Herbert Solomon  
Department of Statistics  
Stanford University  
Stanford, CA 94305

1

Professor Richard L. Dykstra  
Department of Statistics  
University of Missouri  
Columbia, Missouri 65201

1

Professor R. L. Disney  
Department of Industrial Engineering  
and Operations Research  
Virginia Polytechnic Institute  
and State University  
Blacksburg, VA 24061

1

Professor Franklin A. Graybill  
Department of Statistics  
Colorado State University  
Fort Collins, CO 80523

1

Dr. D. E. Smith  
Desmatics, Inc.  
P.O. Box 618  
State College, PA 16801

1

Professor J. S. Rustagi  
Department of Statistics  
Ohio State University Research  
Foundation  
Columbus, Ohio 43212

1

	Copies
Professor E. J. Dudewicz Department of Statistics Ohio State University Research Foundation Columbus, Ohio 43212	1
Professor Joseph C. Gardiner Department of Statistics Michigan State University East Lansing, MI 48824	1
Professor Peter J. Huber Department of Statistics Harvard University Cambridge, MA 02318	1
Dr. H. Leon Harter Department of Mathematics Wright State University Dayton, Ohio 45435	1
Professor F. T. Wright Department of Mathematics University of Missouri Rolla, Missouri 65401	1
Professor Tim Robertson Department of Statistics University of Iowa Iowa City, Iowa 52242	1
Professor K. Ruben Gabriel Division of Biostatistics Box 630 University of Rochester Medical Center Rochester, NY 14642	1
Professor J. Neyman Department of Statistics University of California Berkeley, CA 94720	1
Professor William R. Schucany Department of Statistics Southern Methodist University Dallas, Texas 75275	1

	Copies
Professor T. P. Hettmansperger Department of Statistics The Pennsylvania State University University Park, PA 16801	1
Professor Samuel Kotz Department of Management Science and Statistics University of Maryland College Park, MD 20742	1
Professor Gene H. Golub Department of Computer Science Stanford University Stanford, CA 94305	1
Professor P.A.W. Lewis Department of Operations Research Naval Postgraduate School Monterey, CA 93940	

# DISTRIBUTION LIST

No. of Copies

Naval Postgraduate School  
Monterey, CA 93940

Attn: Code 55Mt	1
Code 55As	1
Code 55Bn	1
Code 55Bw	1
Code 55Cu	1
Code 55Ei	1
Code 55Ey	1
Code 55Fo	1
Code 55Gv	1
Code 55Hh	1
Code 55Hk	1
Code 55Hl	1
Code 55Jc	1
Code 55La	1
Code 55Lw	1
Code 55Ls	1
Code 55Mg	1
Code 55Mh	1
Code 55Mu	1
Code 55Mp	1
Code 55Ni	1
Code 55Py	1
Code 55Pk	1
Code 55Re	1
Code 55Rh	1
Code 55Ro	1
Code 55Sy	1
Code 55Su	1
Code 55Ta	1
Code 55Tw	1
Code 55Ty	1
Code 55Ws	1
Code 55Ze	1
L. Ishii	1



DATE  
FILMED  
-8